

# Phylogenomic Evidence for Separate Acquisition of Plastids in Cryptophytes, Haptophytes, and Stramenopiles

Denis Baurain<sup>†,1,2</sup> Henner Brinkmann,<sup>1</sup> Jörn Petersen,<sup>3</sup> Naiara Rodríguez-Ezpeleta,<sup>‡,1</sup> Alexandra Stechmann,<sup>4</sup> Vincent Demoulin,<sup>2</sup> Andrew J. Roger,<sup>4</sup> Gertraud Burger,<sup>1</sup> B. Franz Lang,<sup>1</sup> and Hervé Philippe<sup>\*,1</sup>

<sup>1</sup>Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, Québec, Canada

<sup>2</sup>Algologie, Mycologie et Systématique expérimentale, Département des Sciences de la Vie, Université de Liège, Liège, Belgium

<sup>3</sup>DSMZ—Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Braunschweig, Germany

<sup>4</sup>Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

<sup>†</sup>Present address: Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, Université de Liège, Liège, Belgium.

<sup>‡</sup>Present address: Functional Genomics Unit, CIC bioGUNE, Derio, Bizkaia, Spain.

\*Corresponding author: E-mail: herve.philippe@umontreal.ca.

Associate editor: Martin Embley

## Abstract

According to the chromalveolate hypothesis (Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol* 46:347–366), the four eukaryotic groups with chlorophyll *c*-containing plastids originate from a single photosynthetic ancestor, which acquired its plastids by secondary endosymbiosis with a red alga. So far, molecular phylogenies have failed to either support or disprove this view. Here, we devise a phylogenomic falsification of the chromalveolate hypothesis that estimates signal strength across the three genomic compartments: If the four chlorophyll *c*-containing lineages indeed derive from a single photosynthetic ancestor, then similar amounts of plastid, mitochondrial, and nuclear sequences should allow to recover their monophyly. Our results refute this prediction, with statistical support levels too different to be explained by evolutionary rate variation, phylogenetic artifacts, or endosymbiotic gene transfer. Therefore, we reject the chromalveolate hypothesis as falsified in favor of more complex evolutionary scenarios involving multiple higher order eukaryote–eukaryote endosymbioses.

**Key words:** eukaryote–eukaryote endosymbioses, chromalveolate hypothesis, phylogenomic falsification, variable length bootstrap, multigene analysis.

## Introduction

Bacterial endosymbiosis has been a major evolutionary force in shaping eukaryotic cells as we know them today. In an ancient event, an alphaproteobacterium gave rise to mitochondria, which house respiration and oxidative phosphorylation along with a variety of other functions (Embley and Martin 2006). A subsequent endosymbiosis with a cyanobacterium resulted in plastids capable of photosynthesis. These “primary” plastids occur in Plantae (Archaeplastida), that is, green algae and land plants, red algae, and glaucophytes (Palmer 2003; Reyes-Prieto et al. 2007; Gould et al. 2008). More recently in evolutionary history, nonphotosynthetic eukaryotes came across another way of acquiring plastids: engulfing a photosynthetic eukaryote instead of a photosynthetic bacterium. When the symbiont is a member of Plantae, the event is termed “secondary” endosymbiosis, while “tertiary” (“quaternary” etc.) endosymbiosis designates engulfment of a symbiont that is already the product of a preceding eukaryote–eukaryote endosymbiosis (EEE). Evidence for EEEs comes from the presence of three or four membranes surrounding the plastid, which requires targeting of nucleus-encoded plastid proteins with multipartite pre-

sequences (Cavalier-Smith 2003), and from less reduced states, where the endosymbiont retains a vestigial nucleus (nucleomorph) (Ludwig and Gibbs 1987).

Cryptophytes, alveolates, stramenopiles (heterokonts), and haptophytes (in the following collectively referred to as CASH) are four diverse and ecologically important eukaryotic lineages that include both photosynthetic and nonphotosynthetic taxa. CASH plastids likely arose from a single initial event of secondary endosymbiosis with a red alga (Palmer 2003; Reyes-Prieto et al. 2007; Gould et al. 2008) because 1) all photosynthetic species of CASH have chlorophyll *c*, which is absent from all other algae (including reds), and 2) in phylogenies based on plastid-encoded genes (Yoon et al. 2002), as well as on certain nucleus-encoded proteins involved in plastid function, for example, plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (Fast et al. 2001), class II fructose-1,6-bisphosphate aldolase (FBA-II) (Patron et al. 2004), and phosphoribulokinase (PRK) (Petersen et al. 2006), a monophyletic group of CASH lineages is recovered. Yet, it has been controversial whether or not CASH plastids (and nucleus-encoded gene products functioning in plastids) were inherited strictly vertically, as postulated in the

popular chromalveolate hypothesis (Cavalier-Smith 1999, 2003; Keeling 2009) (fig. 1A), or spread horizontally, as suggested by “serial” or “multiple EEE” hypotheses (Cavalier-Smith et al. 1994; Sanchez-Puerta et al. 2007; Archibald 2009) (fig. 1B).

By positing a single endosymbiosis at the origin of CASH lineages, the chromalveolate hypothesis is the most parsimonious, though implying two debatable assumptions: First, that EEE events are exceedingly rare, and second, that the numerous plastid-lacking CASH species have all lost their plastids secondarily and independently. In the competing serial hypotheses, chlorophyll *c*-containing plastids also originate by secondary endosymbiosis of a red alga within a eukaryotic host; however, plastids subsequently spread to distantly related CASH hosts via independent higher order EEEs, leading to evolutionary chimeras resembling Russian dolls (Cavalier-Smith et al. 1994; Boudry 2005; Sanchez-Puerta et al. 2007; Teich et al. 2007; Archibald 2009). Accordingly, the observed monophyly of plastid-targeted nuclear gene products, such as GAPDH, FBA-II, and PRK, would not be due to direct descent. Instead, driven by coevolution with the serially transferred plastids, these genes would relocate from the vanishing nucleus of the photosynthetic endosymbiont (or nucleomorph) into the nucleus of CASH hosts (Petersen et al. 2006; Teich et al. 2007). In recent years, plausibility of serial hypotheses has increased with confirmed examples of independent higher order EEEs in dinoflagellates (Hackett et al. 2004).

At present, both hypotheses are equally short of phylogenetic evidence. A key prediction of the chromalveolate hypothesis, that is, the monophyly of CASH lineages, has been extensively tested. Yet, results have been so far inconclusive because cryptophytes and haptophytes cannot be robustly located by phylogenomic analyses of nuclear genes (Burki et al. 2007, 2008, 2009; Hackett et al. 2007; Patron et al. 2007). Moreover, when reduced to the concept of a single endosymbiosis at the base of CASH lineages, the chromalveolate hypothesis can accommodate phylogenetic counterevidence by merely postulating additional plastid losses (Keeling 2009). A point in case are the recent studies indicating that alveolates and stramenopiles are more related to Rhizaria than to cryptophytes and haptophytes (Burki et al. 2007, 2008, 2009; Hackett et al. 2007), leading to the addition of this diverse supergroup to chromalveolates even if most Rhizaria are nonphotosynthetic—thus do not share any of the characters originally used to unite CASH lineages (Cavalier-Smith 1999, 2003). Curiously, the few photosynthetic species within this group harbor plastids of green algal origin (Palmer 2003; Reyes-Prieto et al. 2007; Gould et al. 2008).

Here, we present a phylogenomic falsification (Popper 1959) of the chromalveolate hypothesis that does not require full resolution of the eukaryotic tree. The central idea is that the signal strength for CASH monophyly should be comparable across plastid, mitochondrial, and nuclear genes owing to a common history of the three genomes (Patron et al. 2007). To this end, we devise a taxon sampling for which the chromalveolate hypothesis predicts the same

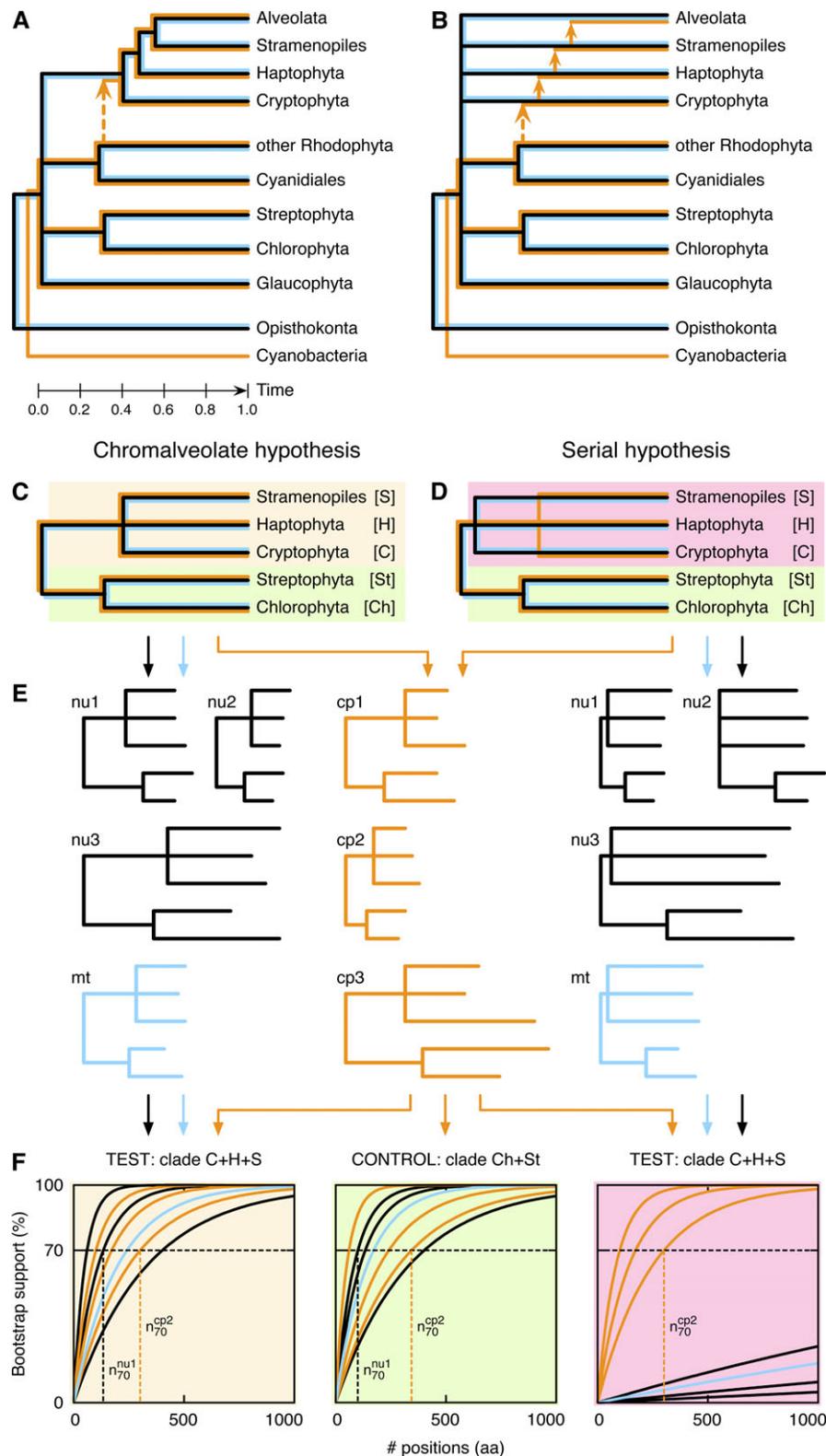
tree for the three genomic compartments and we use the well-accepted plastid history for signal calibration. Further, we develop a protocol ensuring that neither heterogeneities of the evolutionary process nor limitations of inference methods nor endosymbiotic gene transfer (EGT) can impede deciding between the two competing scenarios.

### Protocol

Although the chromalveolate hypothesis states that CASH plastid, mitochondrion, and nucleus all share the same evolutionary history after the red algal endosymbiosis event, the overall history is not shared by the three genomes, with CASH plastids emerging from within red algae (fig. 1A, orange lines). To allow direct comparison of the plastid tree with mitochondrial and nuclear trees, red algae have to be removed. Hence, when using an ingroup composed of CASH lineages, green plants, and glaucophytes, all three genomes should yield a single tree featuring a long branch at the base of CASH lineages (fig. 1C), which corresponds to a strong phylogenetic signal for their monophyly. In contrast, serial hypotheses entail that mitochondrial and nuclear trees should be identical and both differ from the plastid tree, with CASH lineages diverging deeply, either barely supported by a short basal branch or displaying para- or polyphyletic relationships (blue and black lines in fig. 1B and D). Consequently, the phylogenetic signal for the monophyly of CASH lineages should be weak (or nonexistent) with mitochondrial and nuclear genes, whereas with plastid genes, it should be as strong as in the chromalveolate hypothesis (fig. 1C and D, orange lines; see also [supplementary fig. S1](#), Supplementary Material online, for a stepwise illustration of this argument and [supplementary figs. S2 and S3](#), Supplementary Material online, for the relative dating of plastid genomes that provides branch lengths for fig. 1A–D).

In the absence of red algae and assuming equal evolutionary rates and perfect phylogenetic methods, the chromalveolate hypothesis is the only one to predict that the monophyly of CASH lineages should be recovered with an equally small number of positions from plastid, mitochondrial, and nuclear genomes. To compare signal strength across these genomes, we use  $n_{70}$ , defined as the number of positions needed to reach a statistical support greater than 70% using a variable length bootstrap (VLB) strategy. Considering that genes evolve differently in each of the three genomes (owing to, e.g., specific modes of mutation and selection or unequal effective population sizes) and that substitution rates are uneven across genes and over time, along with the usual shortcomings of tree inference methods (Sanderson and Shaffer 2002; Philippe et al. 2005), uncertainties in measurements of signal strength are expected. To be able to decide between chromalveolate and serial hypotheses, it is thus critical to ensure that the range of  $n_{70}$  estimates is narrow in regard to the amount of genomic sequences available for our falsifying experiment (Popper 1959).

To explore the real width of this range, we developed the protocol depicted in [figure 1](#). First, we jointly use



**FIG. 1.** Deciding between two competing hypotheses for the origin and diversification of chlorophyll *c*-containing algae. The evolutionary history of each genomic compartment is shown in distinct colors: plastid (orange), mitochondrion (blue), and nucleus (black). (A) Chromalveolate hypothesis. Except for intra-CASH relationships, branches were scaled by relative dating based on plastid genomes (supplementary figs. S2 and S3, Supplementary Material online). (B) Example of a serial hypothesis scenario (numerous alternatives may be considered). Both chromalveolate and serial hypotheses postulate an initial, single secondary endosymbiosis of a red alga within a eukaryotic host (indicated by dotted arrows), leading to the emergence of the chlorophyll *c*-containing founder. The chromalveolate hypothesis then assumes that this founder alga gives rise to the diverse CASH lineages by vertical descent. In contrast, serial hypotheses invoke multiple subsequent EEEs (indicated by plain arrows), which horizontally spread chlorophyll *c* plastids among otherwise unrelated eukaryotes (C and D).

mitochondrial and nuclear genomes, which share a unique history irrespective of which hypothesis is true (fig. 1A and B), but evolve under specific mutational and selective constraints. Second, to account for the full breadth of evolutionary variation across genes, we split both plastid and nuclear genomes into three functional classes, each displaying distinct evolutionary rates and properties (Philippe et al. 2004; Rodriguez-Ezpeleta, Philippe, et al. 2007). By analyzing these gene partitions separately, we get an insight into the extent of rate variation (fig. 1E; note that this procedure cannot be applied to the mitochondrial genome because of its reduced gene set). Third, we compute  $n_{70}$  estimates (fig. 1F) from trees obtained with two different methods—maximum parsimony (MP) and maximum likelihood (ML). The radically different way in which both approaches extract phylogenetic signal will help to reveal potential tree building artifacts.

Our protocol can be validated using undisputed monophyletic groups as controls. For example, the grouping of chlorophytes and streptophytes into green plants should display similar  $n_{70}$  estimates across all gene partitions (fig. 1F, middle). Concerning the putative monophyly of CASH lineages (here represented by cryptophytes, stramenopiles, and haptophytes), we expect markedly different signal strengths for the two competing hypotheses. If the chromalveolate hypothesis is correct,  $n_{70}$  estimates for mitochondrial and nuclear genomes should be of the same magnitude as those computed for the plastid genome (fig. 1F, left). Alternatively, if  $n_{70}$  estimates are far greater for mitochondrial and nuclear genomes than for the plastid genome, serial hypotheses should be preferred (fig. 1F, right).

## Materials and Methods

### Sequencing

Plastid DNA from *Pavlova lutheri* was purified from a whole-cell lysate (Lang and Burger 2007) and sequenced using a random breakage procedure (Burger et al. 2007). For the nucleus, a cDNA library containing  $1.5 \times 10^6$  recombinant clones was constructed from 5  $\mu$ g polyadenylated messenger RNA using the  $\lambda$ ZAPII synthesis kit (Stratagene) (Petersen et al. 2006). The 5' portion of 1,000 randomly excised cDNA clones was then sequenced.

### Data Assembly

Three large taxon-rich data sets were assembled, each combining protein sequences encoded in one of the three genomic compartments. For the falsifying experiment, plastid

and nuclear gene collections were both split into three functional classes to yield six partitioned data sets (supplementary tables S1–S3, Supplementary Material online). Then, to meet the theoretical constraints depicted in figure 1, the sampling of all data sets was reduced either to 24 species (excluding red algae; table 1) or to 21 species (excluding CASH lineages; table 2), while maximizing similarity of taxon sampling across compartments (supplementary tables S4–S6, Supplementary Material online). A summary of the properties of the 24-species data sets is available in supplementary table S7 (Supplementary Material online). For relative dating and nucleomorph analyses, the taxon-rich plastid and nuclear data sets were used, respectively.

Mitochondrial sequences were retrieved from GOBASE (<http://gobase.bcm.umontreal.ca/>), while plastid and nuclear alignments from Rodriguez-Ezpeleta et al. (2005) were updated with new sequences from GenBank (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein>), dbEST (<http://www.ncbi.nlm.nih.gov/dbEST/>), and the Trace Archive (<http://www.ncbi.nlm.nih.gov/Traces/>) at the National Center for Biotechnology Information.

Single-gene alignments were assembled using new features of the program ED from the MUST software package (Philippe 1993). Ambiguously aligned regions were detected and removed with GBLOCKS (Castresana 2000); this automated selection was slightly refined by hand using NET (also from MUST).

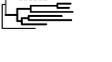
Concatenations of single-gene alignments into supermatrices were carried out with SCAFOS (Roure et al. 2007). When multiple orthologous sequences were available for a particular operational taxonomic unit (OTU), SCAFOS helped to select the slowest evolving sequence as determined from ML distances computed under a WAG + F model with TREE-PUZZLE (Schmidt et al. 2002). To minimize the amount of missing data, SCAFOS was allowed to create chimerical OTUs by merging partial sequences from closely related species (supplementary table S8, Supplementary Material online) when full-length sequences were not available.

To reduce the computational burden associated with bootstrap analyses of the large nuclear data set, amino acid positions missing in  $\geq 30\%$  OTUs were discarded prior to phylogenetic inference, thus resulting in a final supermatrix of 15,392 positions (instead of 19,933 positions for the raw 108-gene concatenation). For single-gene analyses, aimed at uncovering EGT, protein alignments were first cleared of sequences with more than 50% missing positions, to minimize stochastic error due to partial sequences.

←

Phylogenetic predictions of these two hypotheses when red algae, which constitute a special case as the plastid donor to CASH lineages, are not included (glaucophytes and the outgroup are omitted from the drawing for simplicity; alveolates are not considered owing to their highly reduced plastid genomes). The control clade (green plants) is shaded in green, whereas the test clade (CASH lineages) is shaded in yellow or pink for chromalveolate (C) and serial hypotheses (D), respectively. (E) Plastid (cp1–3), mitochondrial (mt), and nuclear (nu1–3) trees expected from different data sets extracted from the three genomic compartments under chromalveolate (left, middle) or serial hypotheses (middle, right). (F) Plots of VLBs and computation of  $n_{70}$  values to test the monophyly of green plants (middle) or of CASH lineages (left, chromalveolate hypothesis; right, serial hypothesis). Plot backgrounds match shadings in (C and D). See text and supplementary figure S1 (Supplementary Material online) for details.

**Table 1.** Signal for the Monophyly of Green Plants and CASH Lineages. ML Trees and VLB Plots Using the WAG +  $\Gamma_4$  Model are Shown at the Same Scale, with Glaucophytes as Thin Lines, Green Plants (Ch + St) Dashed Lines, and CASH Lineages (C + S + H) Thick Lines. The Same 24 Species (or close relatives) were Included in All Data Sets, Except for Outgroups (supplementary tables S4–S6, Supplementary Material online).

	# Positions	# Subs/Site	ML Trees	ML VLBs		$n_{70}$ Values		
				Ch + St	C + S + H	Ch + St	C + S + H	
<b>Plastid</b>								
Polymerase	1,911	4.79				365 531	1,802 3,695	ML MP
Photosynthesis	5,717	1.63				101 159	95 130	ML MP
Ribosome	2,074	3.56				208 345	172 246	ML MP
Polymerase + Photosynthesis + Ribosome + Others	10,805	2.62				106 184	134 212	ML MP
<b>Mitochondrion</b>								
Mitochondrion	3,106	3.51				176 300	n.c. n.c.	ML MP
<b>Nucleus</b>								
Proteasome	3,102	2.34				501 972	n.c. n.c.	ML MP
Ribosome	8,697	2.72				130 193	n.c. n.c.	ML MP
Varia	3,980	1.99				305 358	n.c. n.c.	ML MP
Proteasome + Ribosome + Varia	15,392	2.46				184 287	n.c. n.c.	ML MP

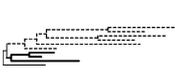
n.c.: not computable.

## Relative Dating

The relative dating was performed on the taxon-rich plastid data set (44 OTUs  $\times$  55 genes) with PHYLOBAYES (<http://www.phylobayes.org/>), which was also used for phylogenetic inference. This is important as the CAT model (Lartillot and Philippe 2004) is known to be less sensitive to long-branch attraction artifacts (LBA) (Baurain et al. 2007; Lartillot et al. 2007; Philippe et al. 2007; Rodriguez-Ezpeleta, Brinkmann, et al. 2007). First, two independent chains were run on the original data set under the CAT +  $\Gamma_4$  model to check for the convergence of the Markov chain Monte Carlo chain and to infer the topology and the branch lengths. Then, 100 pseudoreplicates were generated using SEQBOOT (Felsenstein 2005) and run for a total number of cycles at least equal to three times the number of cycles required for

the convergence of likelihood values and all other parameters. Trees were collected after the initial burn-in period (400 cycles), and CONSENSE (Felsenstein 2005) was used to compute bootstrap support (BS) values for each branch. To avoid a problem of identifiability in the CAT model (i.e., a constant site can be described either by a rate of 0 or by a profile with a single amino acid), all analyses were performed after removal of constant sites. For the dating step, the clockwise assumption was relaxed using a log-normal autocorrelated model (Kishino et al. 2001). Two independent chains were run in parallel using default priors of PHYLOBAYES. After a burn-in of 500 cycles, relative time estimates were collected from both chains and averaged to provide scaling to figure 1A. For the branch at the base of CASH clade, the two time estimates relative to the last

**Table 2.** Signal for the Monophyly of Green Plants and Red Algae. Red Algae (reds) are Shown as Thick Lines. As for table 1, the Same 21 Species were Included in All Data Sets.

	# Positions	# Subs/Site	ML Trees	ML VLBs		$n_{70}$ Values		
				Ch + St	Reds	Ch + St	Reds	
<b>Plastid</b>								
Polymerase	1,911	3.55				137 235	1,226 4,363	ML MP
Photosynthesis	5,717	1.17				102 164	120 165	ML MP
Ribosome	2,074	2.92				577 1,857	853 2,202	ML MP
Polymerase + Photosynthesis + Ribosome + Others	10,805	1.98				107 184	190 323	ML MP
<b>Mitochondrion</b>								
Mitochondrion	3,106	2.29				174 311	109 182	ML MP
<b>Nucleus</b>								
Proteasome	3,102	1.77				371 703	237 337	ML MP
Ribosome	8,697	2.05				91 125	130 187	ML MP
Varia	3,980	1.46				221 325	128 203	ML MP
Proteasome + Ribosome + Varia	15,392	1.87				110 164	136 238	ML MP

common ancestor of Plantae and CASH lineages were equal to 38.8% and 38.4% (mean = 38.6%), respectively.

### Falsifying Experiment

To estimate the signal across plastid, mitochondrial, and nuclear genomes, we computed trees from pseudoreplicates of variable size (VLBs) (Springer et al. 2001) and collected BS values corresponding to the branches leading to various groups of interest. The  $n_{70}$  value—the number of positions needed to reach a BS  $\geq 70\%$  for a given group—were computed after fitting a simplified monomolecular model to the data, according to a procedure inspired by Lecointre et al. (1994).

For each VLB data set, in-house software was used to generate 1,000 pseudoreplicates of each sample size ( $n_1, n_2, \dots, n_x < N$ ) that were analyzed by MP with PAUP\* (Swofford 2002) or by ML under a WAG +  $\Gamma_4$  model (Yang 1993; Whelan and Goldman 2001) with TREEFINDER (Jobb et al. 2004). CONSENSE (Felsenstein 2005) was used to compute

the consensus of the 1,000 trees obtained for each sample size. Output files were automatically parsed for BS values.

For each group in each data set, the nonlinear regression capabilities of the R package (R-Development-Core-Team 2008) were used to fit a simplified monomolecular model to the empirical data using the formula:  $y = 100(1 - e^{-bx})$ , where  $x$  corresponds to sample size and  $y$  to BS. Once the  $b$  parameter is estimated,  $n_{70}$  can be computed as:

$$n_{70} = \frac{1}{b} \ln\left(1 - \frac{70}{100}\right).$$

Mean numbers of substitutions per position (tables 1 and 2) were computed as the tree length inferred by ML under a WAG +  $\Gamma_4$  model with TREEFINDER (Jobb et al. 2004).

### Nucleomorph Analyses

Bayesian inference with the CAT model (Lartillot and Philippe 2004) on the taxon-rich nuclear data set

including the nucleomorph was carried out exactly as on the taxon-rich plastid data set (see Relative Dating), except that we used a burn-in of 500 cycles. ML analyses under the WAG +  $\Gamma_4$  were performed with both TREEFINDER (Jobb et al. 2004) and PHYML (Hordijk and Gascuel 2005), the latter with SPR moves (subtree pruning and regrafting) to minimize the possibility of getting trapped in local optima. Bayesian inferences aimed at handling heterotachy with a covarion model made use of the parallel version of MrBayes (Ronquist and Huelsenbeck 2003). Each analysis used four metropolis-coupled Markov chain Monte Carlo (MCMCMC) chains starting from a random tree and the program default prior probabilities on model parameters. Bayesian posterior probabilities (PPs) were obtained from the majority rule consensus of the trees sampled after the initial burn-in period determined by the convergence of likelihood values and other relevant parameters across MCMCMC generations. A total of 155,000 MCMCMC generations with sampling every tenth generation were run under the WAG + F +  $\Gamma_4$  (control) model, whereas 200,000 generations were run under the WAG + F +  $\Gamma_4$  + covarion model (Huelsenbeck 2002).

### Estimation of Compositional Heterogeneity

The amino acid composition of the 57 OTUs of the taxon-rich nuclear data set was visualized by assembling a  $20 \times 57$  matrix containing the frequency of each amino acid per OTU using the program NET from the MUST package (Philippe 1993). This matrix was then displayed as a 2D plot in a principal component analysis, as implemented in the SAS package (SAS 1999). Compositional properties of the 24-species VLB data sets (supplementary table S7, Supplementary Material online) were also computed with NET.

### EGT Analyses

SEQBOOT (Felsenstein 2005) was used to generate 100 pseudoreplicates of each taxon-rich nuclear alignment that were analyzed with TREEFINDER (Jobb et al. 2004) under a WAG +  $\Gamma_4$  model. Bipartitions (i.e., branches) from single-gene trees were filtered according to three different support thresholds (BS  $\geq$  50%, 70%, or 90%) to generate a list of weakly, moderately, or strongly supported (i.e., testable) bipartitions, respectively. Testable bipartitions from each gene were compared with those actually present in the concatenated ML reference tree (supplementary fig. S6, Supplementary Material online), and genes that had at least one bipartition supported by a BS  $\geq$  70% conflicting with the reference tree were manually inspected (supplementary table S9, Supplementary Material online).

To select the subset of clearly nontransferred genes, 100 pseudoreplicates of each taxon-rich nuclear alignment were generated and analyzed. Single alignments had been previously cleared of seven fast-evolving OTUs (*Micromonas*, *Ostreococcus*, *Cyanidioschyzon*, the nucleomorph of *Guillardia*, *Blastocystis*, *Paramecium*, and *Tetrahymena*) to avoid LBA artifacts and enhance BS. Support for the monophyly of red algae was determined for each gene by parsing CONSENSE (Felsenstein 2005) output files as in

VLB analyses. Finally, the concatenation of the 21 retained genes (including all OTUs) was analyzed under both CAT +  $\Gamma_4$  and WAG +  $\Gamma_4$  models (supplementary fig. S11, Supplementary Material online), as described for the complete data set (see Nucleomorph Analyses).

## Results and Discussion

To apply our protocol to the comparison of chromalveolate and serial hypotheses, we assembled phylogenomic data sets of gene orthologs from each of the three genomic compartments, with a focus on photosynthetic species. As the strength of the phylogenetic signal is expected to vary with species sampling, we selected virtually identical ingroups for phylogenies to be compared, while Cyanobacteria (for plastid trees) and Opisthokonts (for mitochondrial and nuclear trees) were chosen as close outgroups (supplementary tables S4–S6, Supplementary Material online) to minimize phylogenetic artifacts (Delsuc et al. 2005). The resulting alignments from plastid (55 genes, 10,805 aa positions) and nuclear coding genes (108 genes, 15,392 aa positions) were split into functional classes (supplementary tables S1 and S3, Supplementary Material online). For instance, the plastid data set contains three partitions of genes with distinct evolutionary rates and properties (Rodríguez-Ezpeleta, Philippe, et al. 2007), coding for RNA polymerase subunits (1,911 aa positions), for proteins involved in the photosynthetic apparatus (5,717 aa positions), or for ribosomal proteins (2,074 aa positions). Due to its already small size (13 protein-coding genes, 3,106 aa positions), the mitochondrial data set was left undivided (supplementary table S2, Supplementary Material online).

In a first step, we measured the strength of the phylogenetic signal in our data sets by applying the  $n_{70}$  experiment to a noncontroversial control case, the monophyly of green plants (Viridiplantae = Streptophyta + Chlorophyta). This internal control provides a valuable reference point because this group diverged roughly at the same time as secondary red plastids spread (Yoon et al. 2004). As summarized in table 1, evolutionary properties differ across data sets: The mean number of substitutions per position varies from 1.63 to 4.79, while evolutionary rates are heterogeneous among species (compare, e.g., branch lengths for green plants with RNA polymerase subunits vs. other subsets). Despite these heterogeneities, a few hundred positions are enough to reach 70% BS for the monophyly of green plants (average of  $n_{70}$  values is 332, with individual values varying by functional gene class and phylogenetic methodology from 101 to 972). ML is less demanding than MP (average values of 255 vs. 408), which is not surprising given that ML deals more efficiently with multiple substitutions (Jeffroy et al. 2006). With ML (the method used for all following comparisons), the range of  $n_{70}$  values (from 101 to 501) is small relative to the total number of available positions (3,106 to 15,392). This shows 1) that the monophyly of green plants is easily resolved, 2) that estimation of  $n_{70}$  values is robust, and 3) that inherent

difficulties of phylogenetic inference (e.g., rate variability and other model violations) appear to be negligible here. The results obtained with this control case thus suggest that, applied to CASH lineages, our protocol should allow us to decide between chromalveolate (similar phylogenetic resolution across the three genomic compartments) and serial hypotheses (significant differences between plastid and mitochondrial/nuclear data sets).

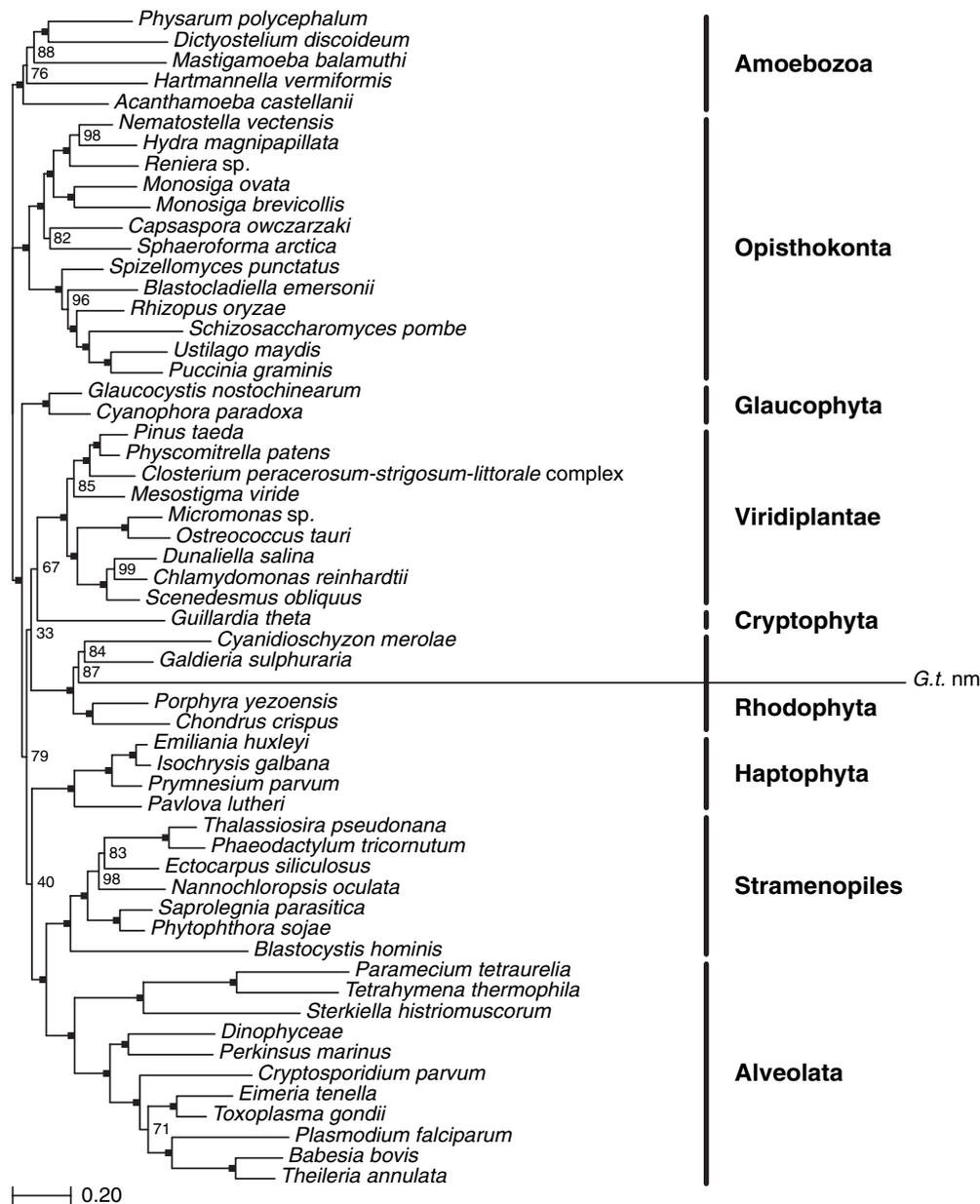
Contrary to the green plant control case, the monophyly of CASH lineages is only recovered by the plastid genome, with a  $n_{70}$  value of 134 (ranging from 95 to 1,802 depending on the functional class; [table 1](#)). Using mitochondrial and nuclear data sets, BS for this group never reaches 20%, even with 15,000 positions—two orders of magnitude higher than required with the plastid data set ([supplementary fig. S4A](#), Supplementary Material online). Such a lack of statistical support for CASH monophyly with mitochondrial and nuclear genomes cannot be explained by rate variation artifacts since branch lengths in mitochondrial and nuclear trees are less heterogeneous than in plastid trees ([table 1](#)). In contrast, the monophyly of CASH lineages is easily recovered with all three plastid partitions, in spite of marked rate differences leading to a 20-fold range of  $n_{70}$  values—though small in regard to the total number of available positions. This includes the extreme case of plastid RNA polymerase subunits ( $n_{70}$  value of 1,802), which evolve at about twice the rate of nuclear genes and which differ by a rate factor of about three between stramenopile and cryptophyte plastids.

Although in [figure 1](#) we do not make assumptions about the relative position of the three Plantae lineages nor about the monophyly of Plantae themselves, the length of the branch leading to CASH lineages, on which our falsification of the chromalveolate hypothesis is conditional, is affected by these relationships. This length depends mainly on the length of the branch at the base of red algae ([supplementary fig. S1C](#), branch 1, Supplementary Material online) and to a lesser extent on the depth (within red algae) of the secondary endosymbiosis event (branch 2) and of the last ancestor of extant CASH plastids (branch 3). In contrast, the contribution of the short internal branches connecting the major eukaryotic lineages is likely to be minor. Yet, the uncertain position of glaucophytes ([Deschamps and Moreira 2009](#)) and/or the hypothetical inclusion of some (or all) CASH lineages within paraphyletic Plantae ([Nozaki et al. 2007](#); [Stiller 2007](#)) might differentially influence the relative length of the branch at the base of CASH lineages across the three genomic compartments. To account for this, VLB analyses were also performed after further removal of either glaucophytes ([supplementary fig. S4B](#), Supplementary Material online) or green plants ([supplementary fig. S4C](#), Supplementary Material online), thus leaving only one Plantae lineage in addition to CASH lineages and outgroups. With these two alternative samplings, results were highly similar to those of the original experiment, which confirms that the short internal branches connecting the major eukaryotic lineages are negligible with respect to the long branches at the base or within red algae ([supplementary fig. S1](#), Supplementary Material online). In

summary, the easy recovery of CASH monophyly with the plastid genome compared with its complete lack of support with mitochondrial and nuclear genomes ([table 1](#) and [supplementary fig. S4](#), Supplementary Material online) is incompatible with the chromalveolate hypothesis that predicts that the monophyly of CASH lineages should be recovered with a similar number of positions, whether sampled from plastid, mitochondrial, or nuclear genes (just as in the green plant control).

To ensure that the easy recovery of green plant monophyly is not a special case, we applied the same protocol to a second noncontroversial group of organisms: the red algae. As for green plants, red algal monophyly should be resolved with a similar amount of positions from all three genomes ([fig. 1A](#)). To do so, CASH species were replaced in the previously used data sets by red algae (*Cyanidioschyzon* plus two non-Cyanidiales species; [supplementary tables S4–S6](#), Supplementary Material online). Since CASH plastids emerge deeply from within red algae ([Yoon et al. 2002](#)), the basal branch of red algae is expected to be slightly shorter than the branch at the base of CASH lineages. Thus, the phylogenetic signal for the monophyly of red algae should be somewhat weaker than the signal for CASH lineages (as measured above; [table 1](#)). With  $n_{70}$  values varying from 120 to 1,226 in plastid data sets and merely from 109 to 237 in mitochondrial and nuclear data sets, [table 2](#) shows that our predictions are perfectly fulfilled. These results demonstrate that our mitochondrial and nuclear data sets have ample power to resolve relationships as ancient as the monophyly of red algae, which lies slightly deeper in the eukaryotic tree than the monophyly postulated by the chromalveolate hypothesis. Such an easy recovery of the monophyly of red algae with mitochondrial and nuclear data in regard to the complete failure to achieve the same goal for CASH lineages ([tables 1](#) and [2](#)) is the most compelling argument against the chromalveolate hypothesis and a strong call for serial hypotheses.

To exclude the possibility that the lack of support for CASH monophyly would be due to a combination of phylogenetic artifacts (e.g., compositional bias coupled to unequal evolutionary rates) ([Rodriguez-Ezpeleta, Brinkmann, et al. 2007](#)), we tested the phylogenetic affinity of a fast-evolving genome: the red alga-derived nucleomorph of the cryptophyte *Guillardia theta* ([Douglas et al. 2001](#)). Nucleomorph genes are characterized not only by evolutionary rates approximately three to eight times higher than nuclear genes of regular red algae and of CASH lineages but also by a strongly biased amino acid composition due to high A + T content ([supplementary fig. S5](#), Supplementary Material online). We assembled an extended nuclear data set including red algae, the cryptophyte nucleomorph, as well as various nonphotosynthetic lineages and species ([supplementary table S6](#), Supplementary Material online). Note that the nucleomorph and the plastid of *Guillardia* were acquired in the same endosymbiotic event, thus implying equal phylogenetic depth. We reason that a successful clustering of the very fast-evolving nucleomorph with red algae would rule out artifacts in resolving



**Fig. 2.** Nuclear phylogeny of CASH species. Bayesian tree based on 15,392 unambiguously aligned amino acid positions from 108 concatenated proteins, obtained using the CAT +  $T_4$  model. While monophyly of all lineages is highly supported, most intergroup relationships remain unresolved (except Opisthokonts), including the monophyly of CASH lineages (BS = 6%). Yet, the extremely fast-evolving nucleomorph of the cryptophyte *Guillardia theta* (G.t. nm) clusters with red algae with 100% BS (at 1,000 aa positions, BS = ~60%; data not shown). A square symbolizes 100% BS; lower values are given. The scale bar indicates the number of substitutions per site.

the slowly evolving and compositionally less-biased CASH lineages. As shown in [figure 2](#), a Bayesian analysis of this extended nuclear data set with the CAT model yields 100% BS for the grouping of the cryptophyte nucleomorph with red algae, while providing no support for CASH monophyly. With the standard WAG model ([supplementary fig. S6](#), Supplementary Material online) or a covarion model handling heterotachy ([supplementary fig. S7](#), Supplementary Material online), similar results are obtained, though with unsupported topological differences. Therefore, except for unknown phylogenetic artifacts and rate variations ([supplementary fig. S8](#), Supplementary Material online),

the divergence of CASH lineages must be considerably deeper than the diversification of red algae, a conclusion incompatible with the chromalveolate hypothesis.

Conflicting signals due to paralogous or horizontally transferred genes are known to cause phylogenetic resolution to decrease, or even to alter tree topologies (Doolittle 1999). In organelles, gene transfer and duplication has been documented in only few instances (Bergthorsson et al. 2003; Rice and Palmer 2006) and to an extent that is insufficient to account for discrepancies between plastid versus mitochondrial phylogenies. In contrast, paralogs frequently occur in nuclear genomes but have been

avoided here by applying stringent phylogenetic selection procedures (Delsuc et al. 2005). This leaves us with detecting artifacts caused by potential EGT (the transfer of genes from an endosymbiont to its host), following EEE events (Stiller 2007; Lane and Archibald 2008). To recognize EGT in the nuclear data set, we applied two protocols. First, we directly analyzed how many single-gene phylogenies diverge from the concatenated phylogeny ( $BS \geq 70\%$ ; using [supplementary fig. S6](#), Supplementary Material online, as the reference tree) and whether these can be explained by EGT. We identified a number of instances in which single-gene phylogenies differ indeed from the concatenated phylogeny ([supplementary fig. S9](#), Supplementary Material online). In most cases (52 events), however, the observed topological rearrangements concern neighboring groups and are best explained by a weak phylogenetic signal; another nine are likely due to LBA artifacts. The four remaining instances have different types of inconsistency that are difficult to interpret but none is consistent with EGT from a red algal nucleomorph to a CASH host cell ([supplementary table S9](#) and [supplementary fig. S10.1–S10.47](#), Supplementary Material online). A second yet indirect check for EGT in our nuclear gene collection is based on the following reasoning. Support for CASH monophyly might be hidden by a conflicting signal caused by EGT from the red algal endosymbiont. To identify genes that could not represent red algal-derived transfers, we searched our collection for those recovering the monophyly of red algae (without including any CASH species;  $BS \geq 70\%$ ). In principle, their phylogenetic analysis should provide strong evidence for CASH monophyly. Still, a concatenation of a subset of 21 clearly non-EGT genes yielded unambiguous support for the clustering of the fast-evolving *Guillardia* nucleomorph within red algae but no support at all for the monophyly of CASH lineages ([supplementary fig. S11](#), Supplementary Material online).

Recently, the nuclear genomes of two diatoms (stramenopiles) have been reported to contain more genes of green than of red algal origin (Moustafa et al. 2009). However, none of our genes (except maybe *rps26*; see [supplementary fig. S10.2](#), Supplementary Material online) appear to contain CASH sequences of green algal origin. Moreover, gene transfer from green algae to CASH species would be expected to equally reduce the support for the monophyly of green plants and CASH lineages; yet [table 1](#) shows that green plants are recovered with a few hundred positions, whereas CASH monophyly is not recovered with 15,000 positions, which further excludes contamination of our collection by EGT.

## Conclusion

In phylogenetics, the lack of resolution is usually dismissed as a nonsignificant result. This reasoning cannot be applied here because easy recovery of the monophyly of green plants and red algae (with a few hundred amino acid positions) is at odds with the complete lack of support for the monophyly of CASH lineages, thus refuting the predictions of the chromalveolate hypothesis. We provide detailed evidence that

this failure cannot be explained by rate heterogeneity and other model violations nor by EGT. Albeit our protocol is fine-tuned for the chromalveolate hypothesis, one might apply this approach to other open phylogenetic questions. In theory, it could help deciding between any pair of evolutionary hypotheses with different phylogenetic structures, provided that at least one hypothesis predicts highly supported relationships. However, as our current approach neither constitutes a statistical test nor yields a *P* value, unambiguously different supports between competing hypotheses are required. Such questions include fusion scenarios proposed for the origin of the eukaryotic cell (see also Embley and Martin 2006), for which a strict phylogenetic evaluation has still to be presented (Poole and Penny 2006).

While our protocol was designed to falsify (Popper 1959) the chromalveolate hypothesis, it cannot assess the monophyly of CASH lineages per se. Actually, much larger phylogenomic data sets might eventually demonstrate that CASH host cells are indeed related, though much more distantly than implied by the chromalveolate hypothesis, as some recent studies tend to suggest (e.g., Burki et al. 2009). However, since such a “monophyly” will likely be achieved at the expense of intermixing CASH lineages with ever more nonphotosynthetic organisms (telonemids and centrohelids, in addition to the aforementioned Rhizaria), it rather constitutes counterevidence for the chromalveolate hypothesis. As a result, this putative assemblage of eukaryotic supergroups, in which photosynthetic organisms represent only a minority, should not be named chromalveolates.

In contradiction to the chromalveolate hypothesis, our results imply that plastid genomes of CASH lineages diverged much more recently than their mitochondrial and nuclear counterparts, which supports serial or multiple EEE hypotheses. There is continuing agreement on an initial secondary endosymbiosis of a red alga within a nonphotosynthetic eukaryotic host, the founder alga eventually evolving chlorophyll *c* plastids. Yet, we posit that this founder alga then engaged in several subsequent higher order EEEs with phylogenetically diverse hosts (in an order still to be resolved), hence giving rise to extant photosynthetic CASH lineages with closely related plastid genomes (Palmer 2003; Reyes-Prieto et al. 2007; Gould et al. 2008) but distantly related mitochondrial and nuclear genomes.

A corollary of accepting serial hypotheses is that complete plastid loss is no longer required to explain the existence of the numerous plastid-lacking CASH lineages, even though recent identification of plastid vestiges convincingly shows the reality of multiple independent losses of photosynthesis (Stelzer et al. 2007; Teles-Grilo et al. 2007; Slamovits and Keeling 2008). Overall, these serial scenarios are more complex and more difficult to decipher. Combination of genome sequencing and cell biology of a multitude of diverse CASH species will be indispensable for inferring more precisely when the original secondary endosymbiosis occurred, which red algal lineage was involved, and how many EEE events led in which order to contemporary chlorophyll *c*-containing algae.

## Supplementary Material

Supplementary figures S1–S11 and tables S1–S9 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

## Acknowledgments

We thank N. Rodrigue, N. Lartillot, M. Melkonian, D. Moreira, and J.M. Archibald for critical reading of initial versions of the manuscript, as well as C.J. Cox and another anonymous referee for constructive reviews. We gratefully acknowledge the financial support provided by Natural Sciences and Engineering Research Council of Canada; Canadian Research Chair Program; Université de Montréal; and the computational resources provided by Réseau Québécois de Calcul de Haute Performance; Belgian Fonds National de la Recherche Scientifique (FNRS; NIC2). From October 2003 to September 2007, D.B. was a postdoctoral researcher of the FNRS at the Université de Liège (Belgium) and is indebted to the FNRS for the financial support during his stay at the Université de Montréal. N.R.-E. has been supported by the Programa de Becas Predoctorales para la Formación de Investigadores (Government of Basque Country).

## References

- Archibald JM. 2009. The puzzle of plastid evolution. *Curr Biol*. 19:R81–R88.
- Baurain D, Brinkmann H, Philippe H. 2007. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? *Mol Biol Evol*. 24:6–9.
- Bergthorsson U, Adams KL, Thomason B, Palmer JD. 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* 424:197–201.
- Bodil A. 2005. Do plastid-related characters support the chromalveolate hypothesis. *J Phycol*. 41:712–719.
- Burger G, Lavrov DV, Forget L, Lang BF. 2007. Sequencing complete mitochondrial and plastid genomes. *Nat Protoc*. 2:603–614.
- Burki F, Inagaki Y, Brate J, et al. (14 co-authors). 2009. Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, Telonemia and Centroheliozoa, are related to photosynthetic chromalveolates. *Genome Biol Evol*. 2009:231–238.
- Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One*. 2:e790.
- Burki F, Shalchian-Tabrizi K, Pawlowski J. 2008. Phylogenomics reveals a new ‘megagroup’ including most photosynthetic eukaryotes. *Biol Lett*. 4:366–369.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17:540–552.
- Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol*. 46:347–366.
- Cavalier-Smith T. 2003. Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). *Philos Trans R Soc Lond B Biol Sci*. 358:109–133; discussion 133–134.
- Cavalier-Smith T, Allsopp MT, Chao EE. 1994. Chimeric conundra: are nucleomorphs and chromists monophyletic or polyphyletic? *Proc Natl Acad Sci U S A*. 91:11368–11372.
- Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*. 6:361–375.
- Deschamps P, Moreira D. 2009. Signal conflicts in the phylogeny of the primary photosynthetic eukaryotes. *Mol Biol Evol*. 26:2745–2753.
- Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284:2124–2129.
- Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG. 2001. The highly reduced genome of an enslaved algal nucleus. *Nature* 410:1091–1096.
- Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. *Nature* 440:623–630.
- Fast NM, Kissinger JC, Roos DS, Keeling PJ. 2001. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol Biol Evol*. 18:418–426.
- Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package). Version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.
- Gould SB, Waller RF, McFadden GI. 2008. Plastid evolution. *Annu Rev Plant Biol*. 59:491–517.
- Hackett JD, Anderson DM, Erdner DL, Bhattacharya D. 2004. Dinoflagellates: a remarkable evolutionary experiment. *Am J Bot*. 91:1523–1534.
- Hackett JD, Yoon HS, Li S, Reyes-Prieto A, Rummele SE, Bhattacharya D. 2007. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of ‘Rhizaria’ with chromalveolates. *Mol Biol Evol*. 24:1702–1713.
- Hordijk W, Gascuel O. 2005. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics* 21:4338–4347.
- Huelsenbeck JP. 2002. Testing a covarion model of DNA substitution. *Mol Biol Evol*. 19:698–707.
- Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet*. 22:225–231.
- Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol*. 4:18.
- Keeling PJ. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J Eukaryot Microbiol*. 56:1–8.
- Kishino H, Thorne JL, Bruno WJ. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol Biol Evol*. 18:352–361.
- Lane CE, Archibald JM. 2008. The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends Ecol Evol*. 23:268–275.
- Lang BF, Burger G. 2007. Purification of mitochondrial and plastid DNA. *Nat Protoc*. 2:652–660.
- Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*. 7(1 Suppl):S4.
- Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 21:1095–1109.
- Lecointre G, Philippe H, Van Le HL, Le Guyader H. 1994. How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. *Mol Phylogenet Evol*. 3:292–309.
- Ludwig M, Gibbs SP. 1987. Are the nucleomorphs of cryptomonads and *Chlorarachnion* the vestigial nuclei of eukaryotic endosymbionts. *Ann N Y Acad Sci*. 503:198–211.
- Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* 324:1724–1726.
- Nozaki H, Iseki M, Hasegawa M, Misawa K, Nakada T, Sasaki N, Watanabe M. 2007. Phylogeny of primary photosynthetic eukaryotes as deduced from slowly evolving nuclear genes. *Mol Biol Evol*. 24:1592–1595.
- Palmer JD. 2003. The symbiotic birth and spread of plastids: how many times and whodunit? *J Phycol*. 39:4–12.

- Patron NJ, Inagaki Y, Keeling PJ. 2007. Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages. *Curr Biol*. 17:887–891.
- Patron NJ, Rogers MB, Keeling PJ. 2004. Gene replacement of fructose-1,6-bisphosphate aldolase supports the hypothesis of a single photosynthetic ancestor of chromalveolates. *Eukaryot Cell*. 3:1169–1175.
- Petersen J, Teich R, Brinkmann H, Cerff R. 2006. A “green” phosphoribulokinase in complex algae with red plastids: evidence for a single secondary endosymbiosis leading to haptophytes, cryptophytes, heterokonts, and dinoflagellates. *J Mol Evol*. 62:143–157.
- Philippe H. 1993. MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res*. 21: 5264–5272.
- Philippe H, Brinkmann H, Martinez P, Riutort M, Baguna J. 2007. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. *PLoS One*. 2:e717.
- Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. *Annu Rev Ecol Syst*. 36:541–562.
- Philippe H, Snell EA, Baptiste E, Lopez P, Holland PW, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol*. 21:1740–1752.
- Poole AM, Penny D. 2006. Evaluating hypotheses for the origin of eukaryotes. *BioEssays* 29:74–84.
- Popper KR. 1959. The logic of scientific discovery. New York: Basic Books.
- R-Development-Core-Team. 2008. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
- Reyes-Prieto A, Weber AP, Bhattacharya D. 2007. The origin and establishment of the plastid in algae and plants. *Annu Rev Genet*. 41:147–168.
- Rice DW, Palmer JD. 2006. An exceptional horizontal gene transfer in plastids: gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters. *BMC Biol*. 4:31.
- Rodriguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Löffelhardt W, Bohnert HJ, Philippe H, Lang BF. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr Biol*. 15:1325–1330.
- Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol*. 56:389–399.
- Rodriguez-Ezpeleta N, Philippe H, Brinkmann H, Becker B, Melkonian M. 2007. Phylogenetic analyses of nuclear, mitochondrial and plastid multi-gene datasets support the placement of *Mesostigma* in the Streptophyta. *Mol Biol Evol*. 24:723–731.
- Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.
- Roure B, Rodriguez-Ezpeleta N, Philippe H. 2007. SCAFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics. *BMC Evol Biol*. 7(1 Suppl):S2.
- Sanchez-Puerta MV, Bachvaroff TR, Delwiche CF. 2007. Sorting wheat from chaff in multi-gene analyses of chlorophyll *c*-containing plastids. *Mol Phylogenet Evol*. 44:885–897.
- Sanderson MJ, Shaffer HB. 2002. Troubleshooting molecular phylogenetic analyses. *Annu Rev Ecol Syst*. 33:49–72.
- SAS. 1999. SAS/STAT user's guide. Cary (NC): SAS Institute Inc.
- Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18: 502–504.
- Slamovits CH, Keeling PJ. 2008. Plastid-derived genes in the non-photosynthetic alveolate *Oxyrrhis marina*. *Mol Biol Evol*. 25: 1297–1306.
- Springer MS, DeBry RW, Douady C, Amrine HM, Madsen O, de Jong WW, Stanhope MJ. 2001. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. *Mol Biol Evol*. 18:132–143.
- Stelter K, El-Sayed NM, Seeber F. 2007. The expression of a plant-type ferredoxin redox system provides molecular evidence for a plastid in the early dinoflagellate *Perkinsus marinus*. *Protist*. 158:119–130.
- Stiller JW. 2007. Plastid endosymbiosis, genome evolution and the origin of green plants. *Trends Plant Sci*. 12:391–396.
- Swofford DL. 2002. PAUP\*. Phylogenetic analysis using parsimony (\*and other methods). Version 4. Sunderland (MA): Sinauer Associates Inc.
- Teich R, Zauner S, Baurain D, Brinkmann H, Petersen J. 2007. Origin and distribution of Calvin cycle fructose and sedoheptulose bisphosphatases in Plantae and complex algae: a single secondary origin of complex red plastids and subsequent propagation via tertiary endosymbioses. *Protist*. 158:263–276.
- Teles-Grilo ML, Tato-Costa J, Duarte SM, Maia A, Casal G, Azevedo C. 2007. Is there a plastid in *Perkinsus atlanticus* (Phylum Perkinsozoa)? *Eur J Protistol*. 43:163–167.
- Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*. 18: 691–699.
- Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol*. 10:1396–1401.
- Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol*. 21:809–818.
- Yoon HS, Hackett JD, Pinto G, Bhattacharya D. 2002. The single, ancient origin of chromist plastids. *Proc Natl Acad Sci U S A*. 99:15507–15512.